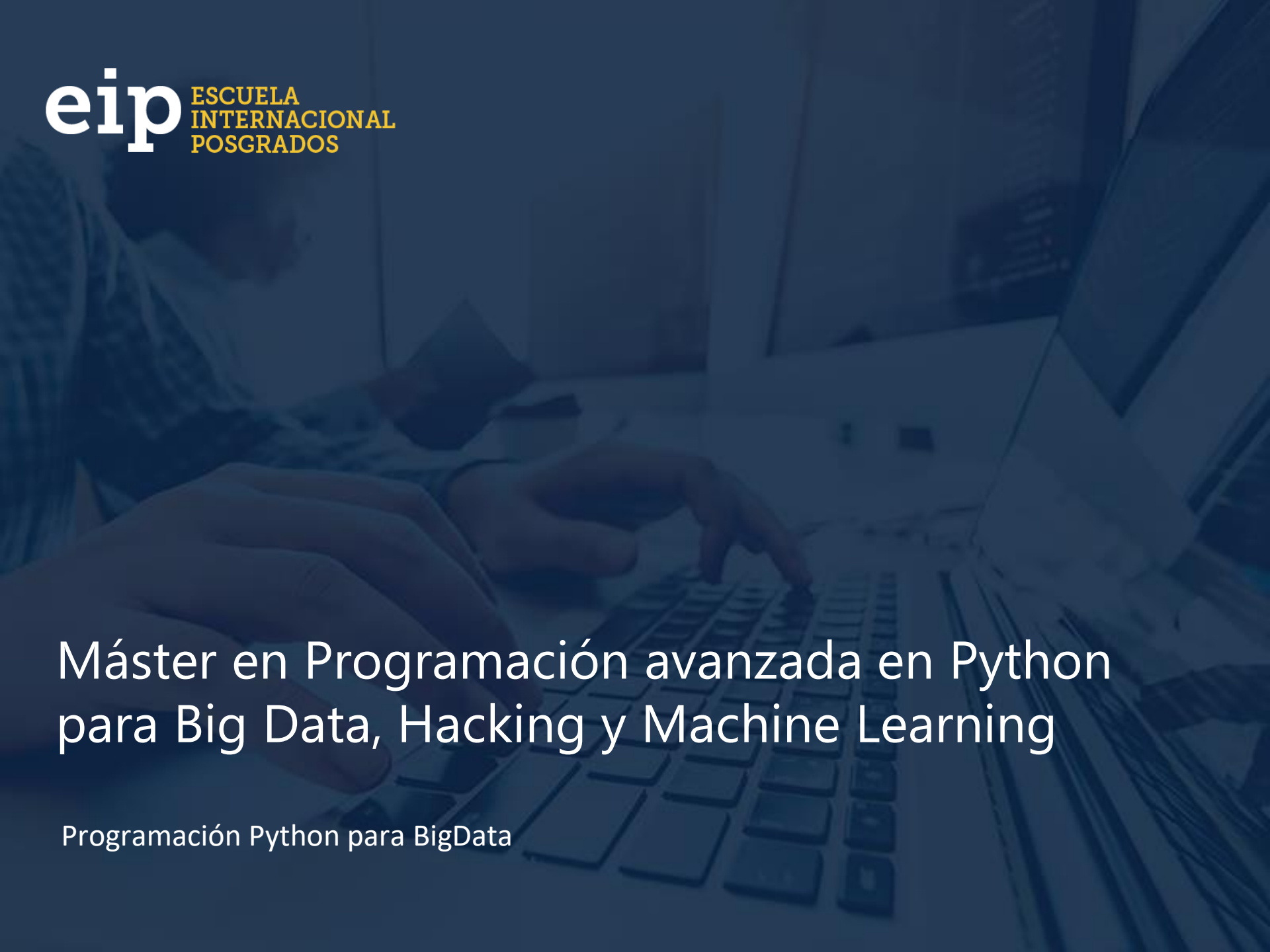


eip ESCUELA INTERNACIONAL POSGRADOS

# Máster en Programación avanzada en Python para Big Data, Hacking y Machine Learning

Programación Python para BigData

# LECCIÓN 06

# VAEX y DASK

# ÍNDICE

# INTRODUCCIÓN

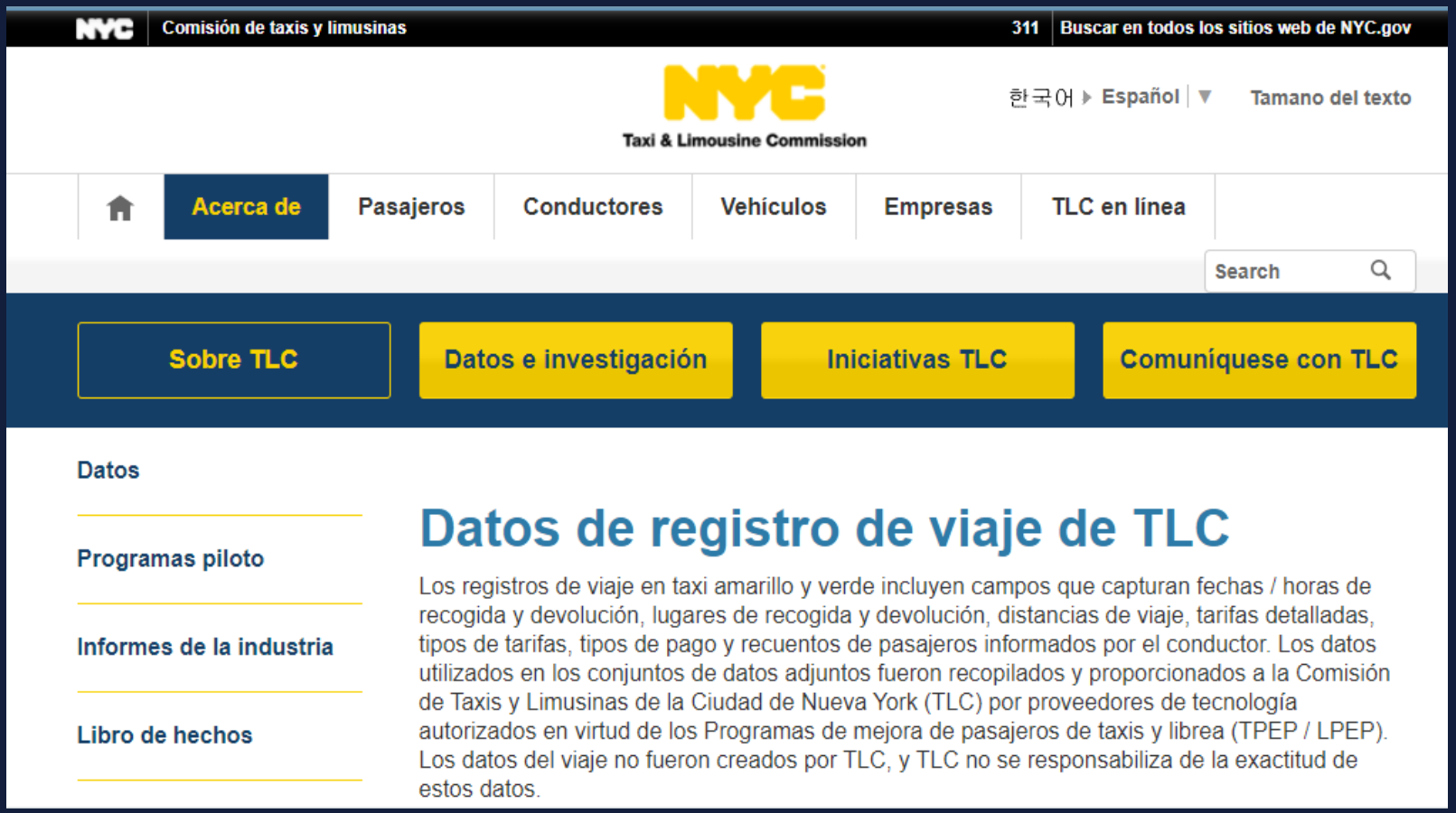
En esta lección aprenderemos a trabajar con un par de herramientas útiles para aquellas situaciones que tenemos un gran número de filas en nuestro dataset.

# OBJETIVOS

**Al finalizar esta lección serás capaz de:**

1. Conocer VAEX como herramienta para manejar grandes volúmenes de datos
2. Conocer DASK como herramienta para manejar grandes volúmenes de datos
3. Medir el tiempo de ejecución de una celda en Jupyter y la repercusión que tiene el ordenador que tienes en el mismo.

# Dataset típico en aprendizaje de Big Data



# Alternativas para manejar grandes volúmenes de datos

- **Vaex**
  Lo veremos en este tema.
  https://vaex.io/docs/datasets.html
- **Dask**
  Lo veremos en este tema
  https://dask.org/
- **Rapids**
  No lo veremos en este manual
  https://rapids.ai/
- **Modin**
  No lo veremos en este manual
  https://modin.readthedocs.io/en/latest/
- **Ray**
  No lo veremos en este manual
  https://ray.io/
- **Koalas**
  No lo veremos en este manual
  https://koalas.readthedocs.io/en/latest/

# CONCLUSIONES

1. VAEX es una gran alternativa cuando queremos hacer Big Data y su principal formato de datos es HDF5

2. DASK es una alternativa a VAEX que en esencia lo que hace es separar el dataset completo en muchos más pequeños pandas dataframes

3. Pandas es una gran herramienta pero quizá no la mejor cuando trabajamos con grandes volúmenes de datos

MUCHAS GRACIAS POR SU ATENCIÓN

eip ESCUELA INTERNACIONAL POSGRADOS

jmpena@grupomainjobs.com

José Manuel Peña
https://www.linkedin.com/in/jos%C3%A9-manuel-pe%C3%B1a-castro-7566b349/

twitter.com/eiposgrados

facebook.com/eiposgrados

instagram.com/eiposgrados